Not Another Review on Computer Vision and Artificial Intelligence in Public Security

A Condensed Primer on Approaches and Techniques

Marcos Vinicius Pinto de Andrade
Software Engineering Department
Cesar School
Recife, Brazil
email:vinivdg@gmail.com

Ana Paula Cavalcanti Furtado
Computing Department
Pernambuco Federal University, UFRPE
Recife, Brazil
email: anapaula.furtado@ufrpe.br

Abstract— **The threats and attacks, perpetrated by criminals and terrorists in many cities around the world have made the use of automated tools for detecting violent acts via video feeds, an invaluable tool to law enforcement authorities. The use of surveillance cameras is widespread, becoming a *de facto* in the security of most cities and makes the use of such content in public security an obvious course of action. The major caveat is the enormous amount of footage that needs to be analyzed, making such tasks not suitable for human operators, and a great candidate for computer vision techniques. The present work aims to bring objective and synthetic information on the subject through a compilation of findings extracted from numerous articles on the subject, serving as a guide for those entering the area: what are the main strategies, approaches, techniques, and features of interest in the area. The paper, in comparison with older but more comprehensive reviews, boasts similar, even though not so comprehensive results, being a valuable starting point for newcomers to this dynamic research area.**

Keywords - artificial intelligence; computer vision; surveillance; public security.

I. INTRODUCTION

Security is among one of the major concerns in modern cities. To address this issue, authorities are using video surveillance on a scale never seen before. This context brings a problem: how to process video streams in a timely manner to avoid damages to people's health or property. The large number of cameras used for surveillance all over the world has created the necessity of streamlining the process of interpreting the large amount of visual data originated from such devices [1]. The obvious choice always leads to some sort of automation, because the amount of data originated from systems with hundreds of cameras will demand an extremely high number of operators, plus coordination and communication strategies in order to work properly, making such setups unpractical in real-world applications [2].

In this context, the advances in computer vision in the past years have made it the technology of choice in any system of automated or intelligent video processing. For this task, a plethora of methods have been developed for processing and analyzing different features or characteristics of video streams. The present research aims to find out what are the most used algorithms, strategies, and tools in computer vision with Artificial Intelligence (AI) for security and surveillance. Since this knowledge area has seen the number of works published grow each year, a study with objective information on how the knowledge in the area is evolving over the years, and what are the best practices used, gains importance serving as a guide for those arriving in the area, bringing guidelines on where to focus the time, resources and energy to make the contribution the most relevant possible.

The first readings in the area showed a myriad of works that, at first sight, seemed very heterogeneous. Further studies showed how many of the approaches are variations of similar algorithms or techniques. The idea for the study is to group all similar approaches, techniques, or algorithms in a way to make clear what are the so-called macro approaches in the area. This work intends to map, in a brief, but insightful way, what techniques of artificial intelligence are being coupled with computer vision techniques for processing security cameras feeds or recordings, aiming at automating violent events detection. Other works approach the same knowledge domain, but the main issue identified is that the search for comprehensiveness has generated works where the big picture, most of the time, is not clear enough for newcomers to the area addressed in this work.

The rest of the paper is structured as follows. Related work is presented in Section II, including the citation of some of the most interesting papers. The methodology is described in Section III and addresses how the papers were selected and the information extracted. The results of the findings are summarized in Section IV, and Section V is a conclusion that brings some observations on the analyzed material.

II. RELATED WORK PANORAMA

An interesting approach using dynamic images, namely a compression of series of video frames into a single bitmap, is presented by Imran et al. in [3]. These images are then fed into MobileNet a Convolutional Neural Network (CNN) for short-term spatio-temporal features extraction. These

 ISBN: 978-1-61208-827-3

features are combined for a representation of the long-term dynamics of the video feed that is analyzed by a Recurrent Neural Network (RNN) that classifies a video content as violent or not. The method also implemented privacy protection layers and is said to have real-time performance capabilities.

Differently from [3], which uses the Optical Flow (OF) of the images, the work by Febin et al. [4] use a Scale-Invariant Feature Transform (SIFT) coupled with a Motion Bound Optical Flow, creating a method that is more robust when dealing with moving camera footage. The work brings results using both Random Forests (RF) and Support Vector Machines (SVM) as classifiers for performance comparison purposes.

In the field of Human Activity Recognition (HAR), [5] bring algorithms based on multi-features processing fed to a CNN for classification. The work claims the approach is reliable in complex real-world scenarios what should open possibilities for use in many areas like smart surveillance for children, elderly, and also uses for entertainment and human-machine interfaces. Interesting, yet simple, work is presented in [6]. A simple layout of a real-world alarm system based on smart surveillance, real word considerations like server topology and other technicalities are worth mentioning.

III. METHODOLOGY CONSIDERATIONS

For the search, the following databases were selected: IEEExplore, Scopus, and Science Direct. The main motivation for this choice is based on the fact that this paper was written during the Covid-19 lockdown and these were the databases that could be accessed with no restrictions from outside the campus.

In a quick summarization, the present research consists of the following stages:

- Paper gathering and selection (including paper search, inclusion and exclusion criteria).
- A quick analysis of the approach used.
- Taxonomy definitions (a database structure with all information classes to be stored and how to do it).
- In-depth analysis of the tools used and/or created on papers.
- Findings compilation.
- Findings uniformization.
- Final synthesis.
- Comparison with similar studies previously selected.

In the selection of the final papers, four works were chosen to draw comparisons with the present work. They were more comprehensive, yet old works, and were used to analyze if the coverage and search quality of the research is acceptable [1]–[6].

The search was conducted initially in an automated way with posterior manual selection phases. Various search strings were tested until the searches began to bring more uniform results. The final search string defined was:

("computer vision" AND surveillance) AND ("computer vision" AND violence) OR ("computer vision" AND harassment)

Four exclusion passes were done after the search. They were based on exclusion criteria applied while title reading and then by abstract examination, and, finally, for the remaining papers, a complete reading was conducted for data extraction and posterior summarization. The first exclusion pass was based on a set of rules defined to maintain uniformity and usefulness of the gathered material, and also to reduce the total number of papers that would be read in full. They are listed below and the final results are depicted in

Figure 1 and **Error! Reference source not found.**. Below are the discriminated criteria.

- Keep a temporal range from 2012 - 2020: When technologies researched began to gain momentum, “the cat experiment” was used as a time mark.
- Eliminate health sciences related material.
- Eliminate other non-security-related material.
- Exclude all material related to Natural Language Processing (NLP).
- Exclude duplicates.
- Exclude non Artificial Intelligence (AI) material.
- Select reviews and surveys, but do not process them.

TABLE I. NUMBER OF PAPERS IN EACH STAGE

	FILTERS PASSES				
DATA BASE	AUTO SEARCH	INC/EXC CRITERIA	TITLE	ABSTRACT	CONTENT
IEEE	23	22	15	13	13
SCIENCE DIRECT	92	85	41	13	8
SCOPUS	35	34	32	18	16
TOTAL	150	141	88	44	37

All papers were gathered in Mendeley [22] for reading and extraction. The resulting data was compiled in an Excel spreadsheet. For databases that did not directly exported CSV files, JabRef [23] was used to do the conversion from RIS or BIB to CSV. The remaining papers were grouped in a single folder for reading and extraction.

The file structure described above permitted the free flow of documents up and down in the folder structure. If one needed to review a discarded document to reconsider a decision, it was instantaneous. All the time, it was possible to have access to all documents in full-text format, which proved to be useful in a small research team configuration, as was the case with the present work.

 ISBN: 978-1-61208-827-3

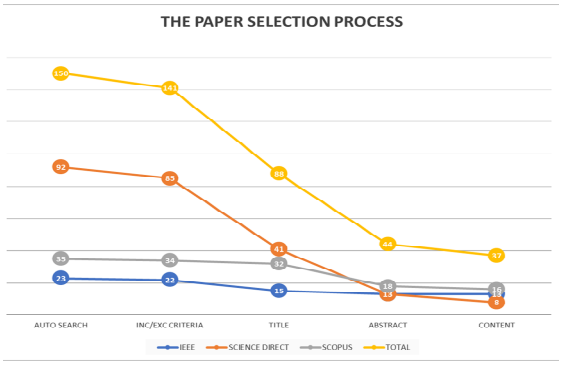

Figure 1. Overview of the selection process, with the amount of paper in each stage of the search/selection.

It was decided to take an approach similar to grounded theory, where a small portion of the material was read to establish a taxonomy of what was important to be extracted to answer the research question. Then, a search for specific pieces of information was conducted inside the papers. The data of interest in the case was:

- What is observed
- Feature identification strategy
- Feature extraction strategy
- Reasoning/classification strategy
- Solution statement by the researcher
- The computational cost of the proposed approach (if present).

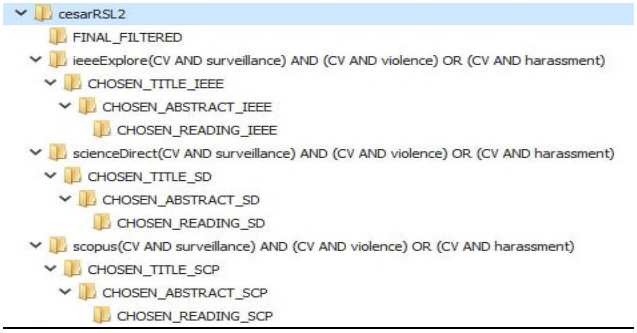

Figure 2. Folder structure inside Mendeley with all the selection stages.

Figure 2 shows a sample of the folder structure created inside Mendeley to process the files. The upper folder always will have all the files downstream, so it was possible to move papers up and down as they were selected or discarded in a given document search phase.

IV. RESULTS

The results can be divided into three main types of researches. First, there were the ones that used some algorithm of computer vision coupled with variants of a Machine Learning (ML) classifier, as in **Error! Reference source not found.**. This approach was found in the vast majority of the works with many variations using modified or enhanced solutions from prior works. A possible explanation may rely on the accuracy and speed of machine learning algorithms like SVM, Random Forests (RF), and their variants. Secondly,

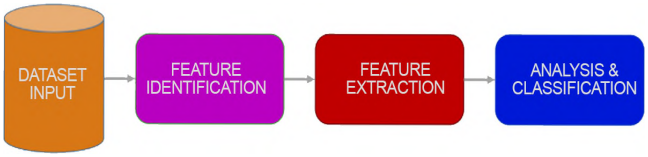

Figure 3. Topology of violence identification systems with Computer Vision and Machine Learning Classifier.

less common initiatives used computer vision algorithms and a kind of neural network for classification ranging from Recursive Neural Networks (RNN) to Deep Learning deployments. At the beginning of this research, there was an *a priori* idea that there will be an emergence of this approach that was not verified in the researched material, which leads to the third macro-approach, depicted in **Error! Reference source not found.**.

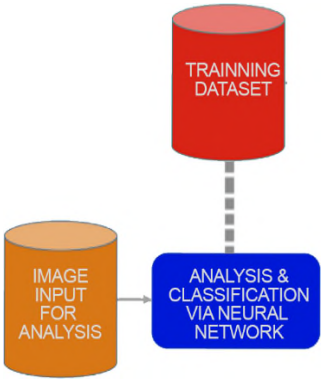

Figure 4. Topology of violence identifications systems with Neural Networks, were found custom trained and pre-trained model used via transfer learning.

The use of pre-trained, custom-trained networks, receiving the direct input of the image feed despite the lower number of occurrences was an approach more consistently identified in the search. The advent of transfer learning is making possible the use of pre-trained networks in many tasks without the burden of training that requires large datasets and more robust computing power. These things are not at easy reach for what was stated in the researched materials.

A. *On what the algorithms process*

Figure 5 summarizes the findings on what is processed in video feeds in the search of violent events. Most of the search features on images for things like edge and gradients

and tracks them in the subsequent frames [3]. The other feature more commonly used was Optical Flow (OF), which is a measure of how the pixels of the image behave over time, which is an indicator of how abrupt the movements in the scene are. These are strong indicators of violent events taking place [7]–[9].

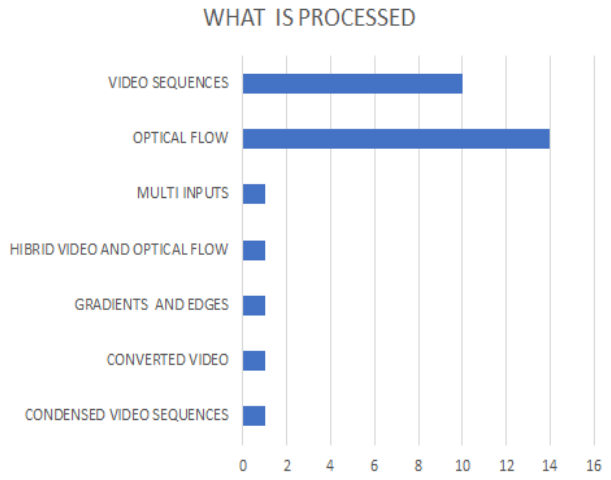

Figure 5. The distribution of findings on what the algorithms look at.

There were ingenious works, like [10], which uses image processing coupled with sensors placed in seats in a public transport vehicle informing when a passenger stands up. Also, it is worth mentioning the use of dynamic images [11][12]. Each approach has its own technicalities, but in general, these are the features the algorithms search within video sequences to classify them as violent or not.

B. *The most used feature identifications and extraction strategies*

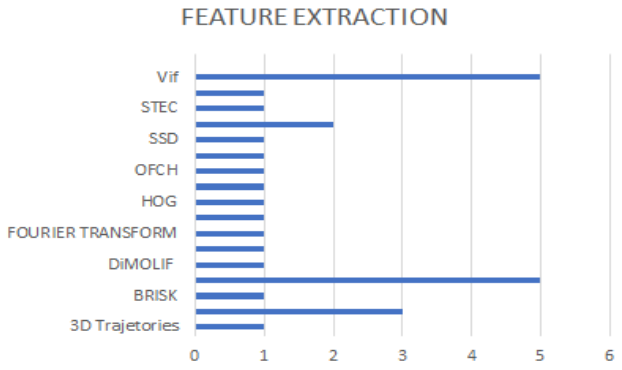

Figure 6. The algorithms and strategies for features identification and extraction of images.

The Violent Flow Descriptor (ViF) uses the variation of the OF magnitude in consecutive frames, being an indicator of abrupt events happening. This approach is present in many of the works. What differs is the subsequent classification model used, which ranges from extremely simple ML models like K Nearest Neighbor (KNN) [13] [14] to deep learning models [15][16].

The big lesson extracted from the material in **Error! Reference source not found.** is that approaches differ in detail, but all the papers used similar strategies with performance improvement modification both in accuracy and computational performance on training and recognition.

C. *The main interpretation and classification techniques used and computational cost issues.*

In the classification area still, ML techniques are prevalent (**Error! Reference source not found.**). SVM and its variations are by far the classifier implemented in almost half the studied solutions. Despite being used for a long time, SVM is a classifier with wide adoption mainly due to the fact that it is non computationally intensive in the training phase and generates models that perform very well near real-time for classifications.

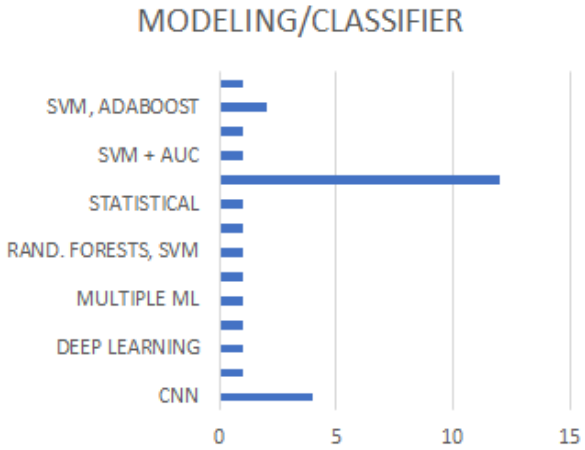

Figure 7. Classification strategies for violent or non-violent definitions.

Our findings raised the question on how are these technologies adoption evolving in time. Are SVM classifiers on their way to retirement? Is deep learning the next big thing in violence detection? For this matter, a short study was done with the evolution of publications on these technologies in the last ten to fifteen years that are presented below (Figure 8) giving a clear panorama on how techniques are evolving. The searches were conducted in Science Direct only, and give a clue on how things are evolving. As can be seen by the graphics, SVM classifiers are still a used choice, probably because of their capabilities and performance On the other hand, deep learning is now beginning to gain momentum on its adoption, being a promising new technology to build new approaches upon.

 ISBN: 978-1-61208-827-3

The use of an optical-flow-based descriptor seems to be reaching a plateau, but it is still relevant.

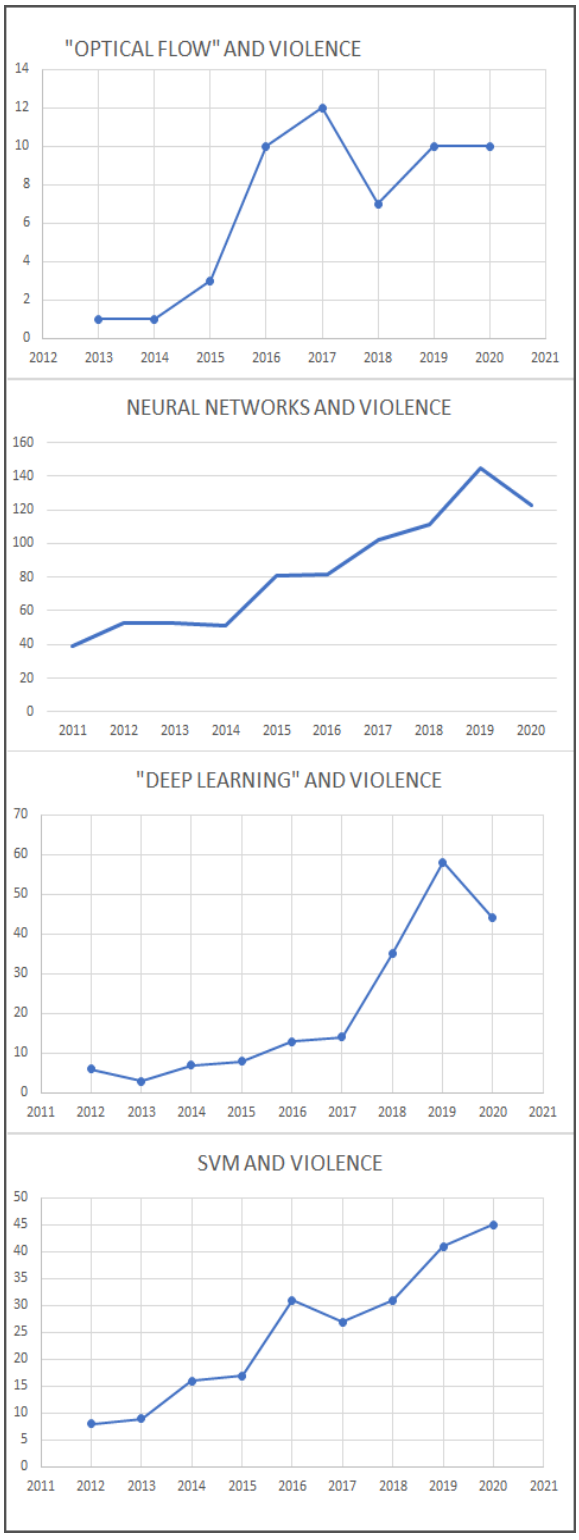

Figure 8: Evolution of the main technologies found in the work and their adoption evolution over time.

D. *Comparison with other reviews*

The scope of the present work was not to exhaust all techniques and approaches, but give a direction on how things are evolving in the academic area on automated surveillance with artificial intelligence and computer vision. Even though not having total comprehension ambitions, the study was able to spot all the main techniques and strategies found in larger and exhaustive studies, like [1]–[6].

V. CONCLUSION

Since the main goal of this research was to give a starting point for those entering the area, some observations on the big picture must be made by the research team. They are as follows:

- Despite being around for quite a while, SVM is still very used.
- The use of deep learning, although being much talked about, appears to be an interesting area to be explored.
- Neural network studies in these areas are being more streamlined by the use of pre-trained networks with good results (transfer learning).
- There was the occurrence of systems that worked in real-time, this being a crucial feature for any surveillance system aiming to prevent violence.
- Some experiments used reenacted scenes as datasets [11], which is an interesting way to supply training data.

TABLE II. ACRONYMS

ViF	Violent Flow Descriptor
STEC	Spatio Temporal Elastic Cuboid
SSD	Single Shot Detection
OFCH	Optical Flow Context Histogram
HOG	Histogram of Optical Gradients
DiMOLIF	Dist. of Magnitude and Orientation of Local Interest Frame
BRISK	Binary Robust Invariant Scalable Key-points
AUC	Area Under Curve
ML	Machine Learning

REFERENCES

[1] M. Ramzan *et al.*, "A Review on State-of-the-Art Violence Detection Techniques," *IEEE Access*, vol. 7, pp. 107560–107575, 2019.

[2] A. C. Nazare Jr. and W. R. Schwartz, "A scalable and flexible framework or smart video surveillance," *Comput. Vis. Image Underst.*, vol. 144, pp. 258–275, 2016.

[3] J. Imran, B. Raman, and A. S. Rajput, "Robust, efficient and privacy-preserving violent activity recognition in videos," in *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, 2020, pp. 2081–2088.

[4] I. P. Febin, K. Jayasree, and P. T. Joy, "Violence detection in videos for an intelligent surveillance system using MoBSIFT and movement filtering algorithm," *Pattern Anal. Appl.*, vol. 23, no. 2, pp. 611–623, May 2020.

[5] A. Jalal, M. Mahmood, and A. S. Hasan, "Multi-features descriptors for Human Activity Tracking and Recognition in Indoor-Outdoor Environments," in *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, 2019, pp. 371–376.

 ISBN: 978-1-61208-827-3

[6] A. Sangeerani Devi, S. Prakash, K. Laavanya, A. Shali, and D. Sathish Kumar, "Violence detection and target finding using computer vision," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 5 Special Issue 3, pp. 235–238, Jul. 2019.

[7] A. Stergiou and R. Poppe, "Analyzing human–human interactions: A survey," *Comput. Vis. Image Underst.*, vol. 188, p. 102799, 2019.

[8] S. Roshan, G. Srivathsan, K. Deepak, and S. Chandrakala, "Chapter 11 - Violence Detection in Automated Video Surveillance: Recent Trends and Comparative Studies," in *Intelligent Data-Centric Systems*, D. Peter, A. H. Alavi, B. Javadi, and S. L. B. T.-T. C. A. in C. C. and I. of T. T. for S. T. S. Fernandes, Eds. Academic Press, 2020, pp. 157–171.

[9] P. Bour, E. Cribelier, and V. Argyriou, "Chapter 14 - Crowd behavior analysis from fixed and moving cameras," in *Computer Vision and Pattern Recognition*, X. Alameda-Pineda, E. Ricci, and N. B. T.-M. B. A. in the W. Sebe, Eds. Academic Press, 2019, pp. 289–322.

[10] A. Boukerche, A. J. Siddiqui, and A. Mammeri, "Automated vehicle detection and classification: Models, methods, and techniques," *ACM Comput. Surv.*, vol. 50, no. 5, 2017.

[11] R. K. Tripathi, A. S. Jalal, and S. C. Agrawal, "Suspicious human activity recognition: a review," *Artif. Intell. Rev.*, vol. 50, no. 2, pp. 283–339, Aug. 2018.

[12] S. Mohammadi, H. Kiani, A. Perina, and V. Murino, "Violence detection in crowded scenes using substantial derivative," in *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2015, pp. 1–6.

[13] E. Y. Fu, H. Va Leong, G. Ngai, and S. Chan, "Automatic Fight Detection in Surveillance Videos," in *Proceedings of the 14th International Conference on Advances in Mobile Computing and Multi Media*, 2016, pp. 225–234.

[14] E. Y. Fu, H. V. Leong, G. Ngai, and S. Chan, "Automatic fight detection in surveillance videos," in *ACM International Conference Proceeding Series*, 2016, pp. 225–234.

[15] M. J. Santofimia *et al.*, "Hierarchical Task Network planning with common-sense reasoning for multiple-people behaviour analysis," *Expert Syst. Appl.*, vol. 69, pp. 118–134, Mar. 2017.

[16] Y. Fan, G. Wen, D. Li, S. Qiu, and M. D. Levine, "Early event detection based on dynamic images of surveillance videos," *J. Vis. Commun. Image Represent.*, vol. 51, pp. 70–75, Feb. 2018.

[17] I. Serrano, O. Deniz, J. L. Espinosa-Aranda, and G. Bueno, "Fight Recognition in Video Using Hough Forests and 2D Convolutional Neural Network," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 4787–4797, Oct. 2018.

[18] X. Xu, S. Gong, and T. M. Hospedales, "Chapter 15 - Zero-Shot Crowd Behavior Recognition," V. Murino, M. Cristani, S. Shah, and S. B. T.-G. and C. B. for C. V. Savarese, Eds. Academic Press, 2017, pp. 341–369.

[19] A. Mumtaz, A. B. Sargano, and Z. Habib, "Violence Detection in Surveillance Videos with Deep Network Using Transfer Learning," in *2018 2nd European Conference on Electrical Engineering and Computer Science (EECS)*, 2018, pp. 558–563.

[20] C. James and D. Nettikadan, "Student Monitoring System for School Bus Using Facial Recognition," in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, 2019, pp. 659–663.

[21] I. Serrano, O. Deniz, G. Bueno, G. Garcia-Hernando, and T.-K. Kim, "Spatio-temporal elastic cuboid trajectories for efficient fight recognition using Hough forests," *Mach. Vis. Appl.*, vol. 29, no. 2, pp. 207–217, Feb. 2018.

[22] Elsevier Mendeley Tool. Available at https://www.mendeley.com/

[23] JabRef Tool. Available at https://www.jabref.org/

 ISBN: 978-1-61208-827-3